GENES VI

Benjamin Lewin

Oxford New York Tokyo
Oxford University Press
1997

shorter sequence is likely to occur—just by chance—a sufficient number of additional times to provide false signals. The minimum length required for unique recognition increases with the size of genome.) The 12 bp sequence need not be contiguous; and, in fact, if a specific number of base pairs separates two constant shorter sequences, their combined length could be less than 12 bp, since the *distance* of separation itself provides a part of the signal (even if the intermediate *sequence* is itself irrelevant).

Attempts to identify the features in DNA that are necessary for RNA polymerase binding started by comparing the sequences of different promoters. Any essential nucleotide sequence should be present in all the promoters. Such a sequence is said to be **conserved.** However, a conserved sequence need not necessarily be conserved at every single position; some variation is permitted. How do we analyze a sequence of DNA to determine whether it is sufficiently conserved to constitute a recognizable signal?

Putative DNA recognition sites can be defined in terms of an idealized sequence that represents the base most often present at each position. A **consensus sequence** is defined by aligning all known examples so as to maximize their homology. For a sequence to be accepted as a consensus, each particular base must be reasonably predominant at its position, and most of the actual examples must be related to the consensus by rather few substitutions, say, no more than 1–2.

More than 100 promoters have been sequenced in *E. coli*, and a striking feature is the *lack of any extensive conservation of sequence* over the 60 bp associated with RNA polymerase. The sequence of much of the binding site is irrelevant. But some short stretches within the promoter are conserved, and they are critical for its function. *Conservation of only very short consensus sequences is a typical feature of regulatory sites (such as promoters) in both prokaryotic and eukaryotic genomes.*

There are four conserved features in a bacterial promoter: the startpoint; the –10 sequence; the –35 sequence; and the distance between the –10 and –35 sequences:

- The startpoint is usually (>90% of the time) a purine. It is common for the startpoint to be the central base in the sequence CAT, but the conservation of this triplet is not great enough to regard it as an obligatory signal.
- Just upstream of the startpoint, a 6 bp region is recognizable in almost all promoters. The center of the hexamer generally is close to 10 bp upstream of the startpoint; the distance varies in known promoters from position –18 to –9. Named for its location, the hexamer is often called the **–10 sequence.** Its consensus is TATAAT, and can be summarized in the form

  $T_{80}\ A_{95}\ T_{45}\ A_{60}\ A_{50}\ T_{96}$

  where the subscript denotes the percent occurrence of the most frequently found base, varying from 45–96%. (A position at which there is no discernible preference for any base would be indicated by N.) If the frequency of occurrence indicates likely importance in binding RNA polymerase, we would expect the initial highly conserved TA and the final almost completely conserved T in the –10 sequence to be the most important bases.
- Another conserved hexamer is centered ~35 bp upstream of the startpoint. This is called the **–35 sequence.** The consensus is TTGACA; in more detailed form, the conservation is

  $T_{82}\ T_{84}\ G_{78}\ A_{65}\ C_{54}\ A_{45}$
- The distance separating the –35 and –10 sites is between 16 and 18 bp in 90% of promoters; in the exceptions, it is as little as 15 or as great as 20 bp. *Although the actual sequence in the intervening region is unimportant, the distance is critical in holding the two sites at the appropriate separation for the geometry of RNA polymerase.*

From data collected on many promoters, we can define the optimal promoter as a sequence consisting of the –35 hexamer, separated by 17 bp from the –10 hexamer, lying 7 bp upstream of the startpoint. The structure of a promoter, showing the permitted range of variation from this optimum, is illustrated in **Figure 11.16.**